DOCUMENT RESUME

ED 079 764 CS 200 657

AUTHOR Effros, Charlotte
TITLE An Experimental Study of the Effects of Guided Revision and Delayed Grades on Writing Proficiency of College Freshmen. Final Report.
INSTITUTION New Haven Univ., West Haven, Conn.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
BUREAU NO BR-2-A-055
PUB DATE Aug 73
GRANT OEG-72-0017(509)
NOTE 28p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *College Freshmen; College Instruction; *Composition (Literary); Conventional Instruction; *Educational Research; Experimental Teaching; Student Teacher Relationship; *Teaching Methods; Writing; Writing Skills

ABSTRACT

Ten sections of Freshman Composition were randomly assigned to either the experimental or control condition. The experimental method, in order to motivate students to revise and rewrite, delayed grades until revisions were completed. The control method used incidental revision with immediate grades. Five instructors each taught one experimental and one control section. Instructional procedures, textbooks, writing assignments, and methods of marking were identical for both groups. Two tests of writing proficiency were applied in a pretest-posttest design. The results showed that for the English Expression Tests the control group was significantly better than the experimental group and that interaction between teacher/class and method was highly significant. For the essay test, the interaction was also highly significant, but there was no significant difference between the methods. (LL)

FILMED FROM BEST AVAILABLE COPY

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: CS
In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view.

Final Report

Project No. 2A055
Grant No. OEG 1-72-0017(509)

Charlotte Effros
University of New Haven
300 Orange Avenue
West Haven, Connecticut 06505

AN EXPERIMENTAL STUDY OF THE EFFECTS OF GUIDED REVISION AND DELAYED GRADES ON WRITING PROFICIENCY OF COLLEGE FRESHMEN

August 1973

U.S. DEPARTMENT OF
HEALTH, EDUCATION AND WELFARE

Field Initiated Studies

National Institute of Education

ABSTRACT

This study was intended to increase our knowledge of the teaching of written composition. Based upon the findings of an earlier study (Buxton, 1958) that careful marking, objective evaluation, and student revision effects a significant improvement in writing skill, this study specifically investigated one of the above-mentioned variables, student revision. In order to motivate students to carefully revise and rewrite, grades were delayed until revisions were completed.

Ten sections of Freshmen Composition were randomly assigned to either the Experimental or Control condition. The Experimental Method was guided revision with delayed grades; the Control method was incidental revision with immediate grades.

Five instructors each taught one experimental and one control section. Instructional procedures, textbooks, writing assignments, and methods of marking were identical for both groups.

Two tests of writing proficiency were applied in a pretest-post-test design. The results, analyzed by means of a two level analysis of covariance, showed that, for the English Expression Tests, the Control group was significantly better than the Experimental group, and that interaction between teacher/class and method was highly significant. For the essay test, the interaction was also highly significant, but there was no significant difference between the methods.

Final Report

Project No. 2A055

Grant No. OEG 1-72-0017(509)

AN EXPERIMENTAL STUDY OF THE EFFECTS OF GUIDED REVISION AND DELAYED GRADES ON WRITING PROFICIENCY OF COLLEGE FRESHMEN

Charlotte Effros
University of New Haven
300 Orange Avenue
West Haven, Connecticut 06505

August, 1973

The research reported herein was performed pursuant to a Grant with the National Institute of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Institute position or policy.

U.S. DEPARTMENT OF
HEALTH, EDUCATION AND WELFARE
Field Initiated Studies
National Institute of Education

Table of Contents

Chapter		Page
I.	Introduction	1
II.	Description of Activities	7
III.	Results	13
IV.	Conclusions	20
V.	Recommendations	21
	Appendix	22
	Bibliography	24

List of Tables

Table		Page
1	Mean Scores for Classes and Groups (Essay Test)	15
2	Analysis of Covariance for a Two-Factor Randomized Design (Essay Test)	16
3	Mean Scores for Classes and Groups (English Expression Tests)	18
4	Analysis of Covariance for a Two-Factor Randomized Design (English Expression Test)	19

I. INTRODUCTION

A. Problem

Teachers of English devote an inordinate amount of time to evaluating (reading, marking and grading) students' themes in teaching written composition. A recent study (Ebel, 1969) estimates that 9-12 hours weekly are devoted to this activity. English Departments of colleges and universities expend more than half of their energies and total semester hours of instruction upon the course in composition usually required of all freshmen (Kitzhaber, 1963). Years of composition research has produced little more than the revelation of what is ineffective. Sherwin, reviewing the research, concludes that "the research overwhelmingly supports the contention that instruction in formal grammar is an ineffective and inefficient way to help students achieve proficiency in writing . . . . that writing does not teach writing; that is, the act of writing alone - the simple increasing of the number of writing opportunities - does not result in a statistically significant improvement in writing skill", (Sherwin, 1969, p. 168) and that the effects of linguistics upon writing seem inconclusive. The purpose of this study is to focus attention upon what the teacher does to and with a student's paper after it has been submitted. Since written expression appears to be so little affected by various methods of instruction prior to the writing experience, it seems reasonable to investigate post-writing activities for clues as to better methods of teaching composition skills.

The traditional method of handling students' themes is to read them, carefully mark all errors, and optimally, write a summary comment which points out the strengths and recommends specific areas for improvement. The papers are then graded and returned to the students. Although the importance of revision is paid lip service, and revision shown to be effective in one major study (Buxton, 1958), such revision is not usually made a requirement of the course, nor are revised themes regraded. If students are to profit from the instruction provided by carefully worded comments and from thoughtfully phrased directions for improvement of a specific theme, it seems reasonable to grade writing after it has been revised. It is the purpose of this research to specifically study the effects of required revision and delayed evaluation upon subsequent writing performance. The Revision Method of handling students' themes was compared to the Traditional Method.

B. Objective

The object of this study was to answer the following questions:

1. Can two markers achieve a high degree of consistency in scoring essays?

2. Is there a statistically significant difference between the writing achievement of college freshmen who have been taught under the Experimental method (Guided Revision and Delayed Grades) and similar students who have been taught under the Control method (Incidental Revision and Immediate Grades)?

3. Is there a statistically significant difference among the classes of each of the five teachers, (each of whom used the Experimental method for one class and the Control method for another class)?

4. Is there a statistically significant interaction between the class/teacher variable and the method variable?

C. Review of Literature

What appears most noteworthy about the vast literature of experimental research on the teaching of composition is that study after study seem to indicate practices which are ineffective. As has already been pointed out, studies of grammar and linguistics have failed to produce significant improvement in composition skill, as measured by performance tests; therefore, this review will focus its attention upon the specific variables under study, marking, grading, and students' revising. The only study discovered by this reviewer to yield statistically significant differences was based on the last variable which is the experimental variable to be investigated herein. Since Buxton's study (1958) has been called the most definitive and the most well designed (Braddock, etal, 1963;) Braddock in Ebel, 1969; Sherwin, 1969), his study will be discussed first and in the greatest detail. Following this, a more cursory review of other related studies will be made. Attention will tend to be focused upon approximately the last decade of research.

Buxton conducted his study at the University of Alberta during one year 1956-1957, and took advantage of the administrative arrangement (all 257 subjects were enrolled in the same courses, for the same number of hours, were evaluated by the same mid-term and final examinations, and were divided into six groups of equal size) to minimize factors external to the experiment.) The experimental purposes were:

> . . . to determine (1) whether or not 'regular practice in writing over a period of seven months (a University of Alberta year) would result in a significant improvement in writing skill and (2) which of the following methods was superior in improving skill: (a) a 'freedom from restraint' or "Writing" method in which no interlinear

> or marginal marks were employed on the students' papers, no grades were noted on the papers; the only remark was a paragraph of generous comment and the students were not advised to correct or revise their papers, and (b) a 'prevision, writing, and revision, or "Revision" method, in which the papers were thoroughly marked, graded, commented upon (adversely, when necessary) in a paragraph at the end of the paper, and discussed by the raters and revised by the students during the 25 to 50 minutes of class time when each paper was returned. (Braddock, etal, 1963, p.58)

Buxton, then, studied the effects of intensive marking, grading, and immediate revision upon the improvement in composition skills by freshmen who were in the Junior Elementary Program, and emergency course to train new teachers for the public schools of Alberta. His research design and rating scale are both worthy of close consideration. The latter, since it will be used in the proposed experiment, will be fully described in Part 2 and a copy appended; suffice it to say for the present that Buxton reports inter-rater reliability coefficients (r) of .91 for the pretest and .88 for the post-test.

In designing his experiment, Buxton took great care to control all extraneous variables. Each of the 257 students was randomly assigned a number, and then the names of students were listed numerically. This list was then ramdomly subdivided into the Control group (86 students), the Writing group (86 students) and the Revision group (85 students). Each of the three groups was treated by a distinctly different method for the purpose of the experiment. The Control group took all their regular subjects, but received no additional writing assignments, as was required of the other two groups. Each student in the Writing and Revision groups took the same courses as the Control students, but wrote additional essays of about 500 words each. These essays, a total of sixteen different themes, were assigned by the instructor of a different course each week, so that each student in the two experimental groups wrote one paper for each course each semester.

For the Writing group, the assignments and markings of papers were characterized by freedom from restraint. Although topics were suggested by the instructor, each student was encouraged to write about his own interest, feelings, and experiences, and could choose his own topic if he so desired. In marking the papers, no interlineal or marginal notation of any type was made, nor were grades given; instead, a three or four sentence commentary, generally praising the paper as much as possible, and offering a suggestion for improving the next paper, was written at the bottom. Papers were returned to students without comment and with no directions for revision or correction. The Revision group was treated with considerably more direction. Topics were specifically assigned, and although students were encouraged to develop their topics in their own way, they were instructed to use a theme sentence, and to develop their material logically and coherently. Preliminary ideas were to be worked into an outline before writing the theme, words and

illustrations were to be chosen with care, and unified paragraphs with logical transitions between them were to be developed. Students were also instructed not to make unsubstantiated statements. In marking, the themes were extensively notated for organization, diction, paragraph unity, overall logic, as well as for errors in spelling, punctuation, and sentence structure; at the bottom of each paper was written a few sentences giving directions for improvement. Two grades were given to each paper, one for content, the other for accuracy of expression. In addition to receiving extensive markings and objective grades, the papers of the Revision group were returned to the students at the beginning of a class meeting for careful correction and revision while the reader circulated from student to student to render whatever assistance seemed needed for accurate revision. However, after revision, the assigned grades were not reconsidered or amended in consideration of the improvements wrought by revisions.

Instructions for the essay pretest and post-test were given on mimeographed sheets. Since the students' backgrounds varied widely, a broad topic was assigned and suggestions made as to different ways in which the topic could be narrowed. The topic of the first of these themes, written during a fifty minute period, was "High Schools," and the topic for the second test was "My Opinion," the latter permitting practically any kind of stand on any subject.

The essay tests were graded by two raters working independently. One was the investigator; the other was the Chairman of the Grade Twelve Essay Marking Department of the Alberta Department of Education. To guide the essays test raters, a score sheet was prepared, developed, and amended by the raters themselves. A feature which may have contributed to the validity of the design is that the score sheet went through several preliminary stages as the raters practiced with it on trial themes. The fact that the raters helped develop the score sheet as they practiced with it suggests that they not only understood it but believed in it and followed its instructions.

After statistical analysis of the results, Buxton found that the three adjusted mean scores (adjusted for differences in pretest scores by means of analysis of covariance) yielded an F-ration of 4.97; which indicated significant differences among the means of the three groups since, at the .01 level of confidence with 2/251 degrees of freedom, an F-ration of 4.695 is significant. Applying individual t-tests to the differences between pairs of adjusted means of post-test, Buxton reports differences significant at the .01 level between the Writing and Control groups, and at the .05 level between the Revision and Writing groups.

The results of Buxton's study appear to support the contention that thorough marking, objective grading, and guided revisions effect improvement in writing; however, it leaves unclear the question of whether it was the marking, the grading, or the revising of papers which caused this improvement.

Sutton and Allen (1964) conducted a study of the effects of writing practice in comparison to no writing practice, and the effects of peer evaluation in comparison to teacher evaluation. One hundred and twelve subjects (45 males and 67 females) were randomly selected from the 1962 freshman class at Stetson University and randomly assigned to two control and three experimental classes. The control classes had no outside writing practice, hence no evaluation. Among the experimental classes, one group wrote for peer evaluation, and another group for teacher evaluation. Statistical analysis of results from two objective tests (English Composition Test of the College Entrance Examination Board and English Expression Tests published by Education Testing Service), summary ratings and rankings of in-class writing exercises yielded no differential effects on written composition which could be attributed either to the frequency of writing practice or to the method of evaluation.

A study by Burton and Arnold (1963) included eight tenth-grade classes with two teachers following four different approaches to intensity of evaluation and frequency of writing. The following approaches were used: (1) infrequent writing with moderate evaluation, (2) frequent writing with moderate evaluation, (3) infrequent writing with intensive evaluation, and (4) frequent writing with intensive evaluation. The study was conducted in two comparable high schools, with a teacher in school teaching four matched groups of students. STEP Essay and STEP Writing Tests were used before and after instruction. Both forms of the essay tests were rated by three experienced raters. A complex factorial design served as the model for the statistical analysis. No main effects or interactive effects were found to be significant at the .05 level. It is interesting to note that the report gives little accurate information as to how many assignments constitute "frequent" or "infrequent" writing, or the nature of the evaluation, whether "intensive" or "moderate." In other words, exactly what were teachers writing on papers? Under the "intensive" conditions, did they merely point out more errors than under the "moderate" condition? In light of the fact that no significant differences were found to be associated with intensity of evaluation or frequency of writing, and that no significant interaction was found, it would seem that further study of these questions is indicated. It seems possible, since intensive evaluation proved ineffective in this study, at least in comparison to moderate, that what the student does with his theme after it has been evaluated may be instrumental in effecting an improvement in his writing proficiency.

In a study of the effects of teachers' comments upon students' motivation, Seidman (1968) used ten classes of high school English students taught by five different teachers. The material, taught similarly by all five teachers, was a sequence of eight composition assignments, which were not described in detail. Within each class, all students were randomly divided into three experimental conditions. Students' papers in Group A received informative and generally reinforcing comments; in Group B, papers received a high proportion of negative, judgmental comments; in Group C, papers received no comments at all. Grades were assigned to all

papers, regardless of the comments, on the basis of the instructional purposes of each assignment, and were checked by the investigator for reasonable consistency between teachers and conditions.

Motivation was inferred by analyzing the number of optional rough drafts and revisions made by all students in the experimental series. Group A students, receiving consistently positive, encouraging comments on their assigned themes, wrote significantly more rough drafts and revised their papers more frequently than students in either Groups B or C. The difference was significant at .025. These findings suggest that positive, non-judgmental comments tend to encourage students to write more, and to revise more carefully. It seems reasonable to assume that when grades are withheld until all revisions are completed and both student and teacher are satisfied with a piece of writing that this increased volume of writing may increase writing proficiency.

In a comparison of the effects of theme correction by teachers with correction by peers, Pierson (1967) studied one hundred fifty-three ninth grade students who were randomly assigned to three experimental and three control classes. Under the experimental conditions, assigned themes were evaluated by the students themselves, working individually and in small groups, by means of a correction chart developed by the researcher. Compositions for the control group were evaluated completely by the teacher. The nature of teacher evaluation, and information as to whether themes were subsequently revised, was not given. The effects of treatment were measured by scores on the STEP Writing Tests given before and after treatment. Analysis of the data revealed no significant difference between the groups as measured by mean score gains on the STEP Writing Tests. Further investigation of the kinds of evaluative activity and the nature of revisions and rewriting that is most effective in causing and improvement in students'writing performance would seem necessary before studies of who should do such evaluation can be fruitful.

In a theoretical article, Zoellner (1969) analyzes the teaching of writing according to the principles of operant conditioning. He points out that the usual practices of red-inking all errors, insisting upon a carefully kept "correction chart," and profusely writing only negative commentary on students' themes clearly contradict the experimental evidence of reinforcement psychology. While most operant conditioning studies in the area of verbal behavior had tended to focus upon verbal-vocal behavior because of the clinician's preoccupation with the interview as a therapeutic technique, "current experimentation in verbal conditioning is nonetheless very suggestive of new angles of attack for the teacher of composition." (p. 294) Since a variety of verbal-vocal responses (including stuttering, expressions of opinion, word classifications such as nouns or adjectives) have been shown to be highly susceptible to reconstruction through the application or verbal reinforcement (expressions of "good," "that's right," etc.), it seems reasonable to assume that written verbal responses are similarly susceptible to modification. It would also seem worthwhile to investigate the

application of Skinner's (1953) technique of shaping behavior by means of differential reinforcement and successive approximation to written verbal behavior. This is the rationale for the experimental method of handling students' themes.

II. DESCRIPTION OF ACTIVITIES

A. Subjects

Subjects for this study were self-assigned to sections of English 113, a required one semester course at the University of New Haven, West Haven, Connecticut. Students who, according to admissions tests and high school English grades, were judged deficient in writing skills, were required to take a non-credit course in Basic Grammar prior to enrollment in English 113. Thus, the lowest ability group were removed from the population. Ten sections of English 113 were chosen from the twenty-two sections in September 1972. The intact classes were then assigned by pairs to each of the five participating teachers to fit their schedules and, within each pair, randomly assigned to either the experimental (Revision) or to the control (Traditional) group.

B. Teachers

The five participating teachers all hold at least a Master's degree. Two are regular, full-time faculty members; the other three have been teaching English 113 at the University of New Haven for a minimum of three years on a part-time basis. Participating teachers met regularly with the investigator, first, prior to the commencement of classes and later at the completion of each unit of work. The assignments and the two methods of evaluating students' themes which were under study were thoroughly discussed at these meetings.

C. Methods of Marking

The methods of marking themes were the same for both experimental and control groups. Interlinear commentary was used to point out mechanics, sentence structure and punctuation, organization of the theme as a whole, and organization and development of paragraphs. The Handbook section of Words and Ideas by Hans Guth was specifically referred to as needed to direct student improvement in subsequent assignments or revision of the current assignment. A summary comment praised everything in the paper that was praiseworthy, and pointed out weaknesses in the theme as a whole, in paragraphing, and in sentence making.

D. Traditional Method

For the control group (Traditional), a grade was assigned and

recorded and the papers returned to the students. No specific attention was given to revisions or rewrites, nor did voluntary revisions receive more than a cursory check. Students were told that they would be expected to build upon their strengths and to avoid their errors in subsequent assignments, and that their grade for the course would depend upon an exhibition of growth in writing skill; in other words, that their grade for the course would be influenced more by their work at the end of the semester than at the beginning in order to provide motivation for careful attention to the teacher's annotation and for effort to improve.

E. Experimental Method

For the experimental group, (Revision), papers were read and annotated in the same manner as for the control group. In order to keep the teachers' method of annotation and procedures in handling papers as nearly equivalent as possible except for the experimental variable (revision), a tentative grade was recorded by the teacher in his grade-book, but not noted on the paper. The summary comment gave specific directions for revising or rewriting that paper. On the day they were returned, (ideally the meeting following submission), explanation and directions for improving themes were given both to the group as a whole, in general terms, and individually in conference in specific terms. Papers were revised or rewritten by the students, and resubmitted at the end of the unit for a single grade, which reflected the students' improvement in writing, as well as the quality of the revised themes.

F. Instructional Procedures

Instructional procedures were as nearly identical as possible for both groups. The course texts, which have been in use for a year and are listed below, were the same; and the writing assignments, which were specifically developed for this project and are described below, were the same. Since all teachers in the experiment taught one section of the control group and one of the experimental, the only systematic variation in treatment was the method in which the papers are handled.

G. Textbooks

Clayes, Stanley A. & Spencer, David G. Contexts for Composition. Second Edition. New York: Appleton-Century-Crofts, 1969.

Guth, Hans P. Words and Ideas. Belmont, California: Wadsworth, 1969. Reading assignments were made from the above texts in order to illustrate and fully explain the writing assignments which form the instructional core of the course.

H. Writing Assignments

Unit I Weeks 1-5 Each theme to be approximately 300 words.

1. Describe a process.
2. Describe a place.
3. Describe a person.
4. Compare and contrast two processes, places, or people.

Unit II Weeks 6-10 Each theme to be approximately 500 words.

5. Define an abstract term.
6. Select an argumentative editorial from a national magazine and argue against it.
7. Same as #6, but using a different editorial. To be written in class.

Each of these assignments was prepared in advance, duplicated, and distributed to all participating students one week prior to the date each assignment was due. Specific topics, material, and focus for each assignment was chosen by the individual student.

In addition to the above listed assignments, all students wrote a fully documented research paper, which was handled in the same manner for both groups and, therefore, is not properly a part of this study.

I. Instrumentation

The results of this experiment was determined by two measures, the Cooperative English Tests, 1960 Edition, published by Educational Testing Service, Princeton, New Jersey, and an essay test, scored by two independent raters using the Buxton scale.

The English Expression Tests of the Cooperative English Tests may be seen as a measure of the student's ability to select an appropriate usage from several alternatives and to recognize an incorrect usage when presented. While the test does not directly measure writing achievement, "evidence suggests that ability to do well on this kind of test is related to writing well in an 'essay' situation." (Manual, 1960, p. 13)

The validity of the English Expression Tests as a measure of writing ability is of crucial importance. The first concern is for content validity, which "is best insured by relying on well-qualified people to construct the test. This was done for the Cooperative English test." (Manual, 1960, p. 13) The predictive validity was studied at the University of Florida, in September, 1958, where Form 1C of the English Expression Tests was administered to 2,449 freshmen. Scores on the English Expression Tests and composite scores of all regular English course tests for the

semester yielded a correlation coefficient of .67.

The reported reliability of the English Expression Tests is based upon alternate form correlation coefficients computed between a pair of forms administered to the same students with a time lapse of a week or less. According to Walter R. Borg, "When more than one type of reliability is computed for a given test, the results of the different types are usually in fairly close agreement." (1963, p. 84) Therefore, the coefficients of equivalency which are quoted may be presumed to be a reasonable estimate of the stability of scores for subjects taking a single form of the English Expression Tests. To determine the reliability, a sample of 20% of the normative population was selected, and each level 1 form was paired with each of the other two forms for level 2. The resulting correlation of .84 for forms 1A and 1B, which was used for this study, is the average of its correlations with the other two forms.

The English Expression Tests would appear to be a reliable and valid objective measure of writing skill. The nature of the items which call for recognition of incorrect usage and the ability to correct errors of diction, mechanics, and usage would seem to provide content validity. Since the experimental method of delayed evaluation and guided revision focuses the students' attention upon correcting errors and improving expression, and the traditional method tries to teach students to avoid errors and write more effectively on subsequent assignments, the results of these methods appear conducive to measurement by this instrument.

The essay test was a 50 minute theme written in class. A single, broad topic entitled "My Opinion", directions for which were duplicated and distributed to all students, was used. Buxton's score sheet and method of scoring were used to rate the essays; the two raters practiced with the score sheets. They clarified and amended as necessary during the summer preceding the experiment until an inter-rater reliability of .85 was achieved. A trial set of student themes from Spring 1972 semester were collected for this purpose.

J. Scoring Procedures

The score sheet, appended, consists of fourteen categories, for which a maximum number of points can be awarded. Each paper received a "basic mark" of 155 points, and then was awarded points for each category. The total number of points are computed, from which deductions were made for errors of spelling, punctuation, usage, grammar, sentence, and form. The theoretical maximum score for each paper was 300 points. Each paper was read at least three times by each rater. During the first reading, the organization of the paper was carefully considered and evaluated,

with examples of effective diction, concreteness, figures of speech, and critical thinking recorded on the back of the score sheet and points awarded for each positive instance. On the second reading, paragraphing, sentence structure, unity, and coherence were considered and evaluated. On the third reading, all errors were tallied under Deductions. When necessary, a fourth reading determined the accuracy of the rating process.

Each paper was identified only by a paper number, which was randomly affixed to the cover of the blue book from a Table of Random Numbers. From each section of students (ranging in size from 15-25 students) ten numbers were randomly selected and a student number assigned to the back of the blue book, and subsequently affixed to the writer's pretest and post-test for the English Expression Tests. This number identified for the investigator the group from which each student came, and there were four papers identified by the same student number (2 Essays and 2 EET's). Paper numbers, however, for essay pre- and post-tests were different for each paper and completely randomized to prevent the readers from identifying the source of a given essay. Blue book covers were filled in by the students, and included such information as teacher, section, and date. These covers were removed by the investigator, the student number written in red on the back page, and the paper number written in blue on both the front and on the back page for identification and collation at the end of the study. As each rater read a paper, the numbers were copied on top of the score sheet. The raters, then, had no way to differentiate between pretest/post-test, or Experimental/Control which could bias their scoring.

The two raters are experienced teachers of composition, each holding a Master's Degree, and were not otherwise involved in the experiment. They worked together during the summer of 1972 to develop the scoring scale, the final version of which appears in the Appendix. In order to maintain consistency, the two raters worked independently on scoring the papers, which were identified only by a student number and a paper number (explained in Chapter II), and then submitted to the investigator, who computed the final score. Periodically, throughout the rating procedure, on the average of every 20th paper, the two raters went over a complete paper, discussed their awarding of points, and compromised any divergence between them, as far as possible. No scores were changed, but subsequent papers were marked according to the compromise.

K. Using the Score Sheet

Even with a highly itemized score sheet, it was necessary for the raters to make many value decisions, which they objectified

and discussed. Illustrative of the decisions they reached are the ones outlined below.

1. The student was awarded two points for each reference to a personal experience, or to a book, or to the statement of an authority. Similarly, any additional piece of supporting evidence was awarded two points, with a maximum of ten points for that category.

2. In deciding upon the marks to be awarded for significance of contents, and for quality of introduction and conclusion, practice papers collected for that purpose were selected, scored, and arranged in rank order of the agreed upon scores. These were subsequently used as models for the scoring of papers in the study.

3. Specific writing qualities such as variety of sentences, transition, figures of speech, startling and appropriate diction were objectively scored by awarding a specific number of points for each instance. For example, any word or phrase which began a paragraph which was obviously intended to connect the material to the preceding paragraph, was awared 2-3 points, depending upon the number of paragraphs in the essay. Points were similarly awarded for diction. Each instance of a word which seemed particularly appropriate to the expressed idea was awarded two points. Following the discussion of sentence structure in Words and Ideas by Hans Guth, two points were awarded for each example of any of the means of sentence variation; for instance, a predicate modifier preceding the subject, a balanced sentence, a sentence containing parallelism, a short effective sentence between longer sentences, etc.

L. Treatment of Errors

Each rater tallied the number of errors on each paper on the back of the score sheets. In order to prevent the markings of one rater from influencing the judgment of the second rater, no marks were made on the papers. To prevent errors from confounding the awarding of positive points, the tallying of errors was the last step made by each rater. The effective use of a word was awarded positive points despite incorrect spelling.

M. The Error Rate

Students were not told how many words to write; therefore, the length of the essays differed considerably. In order to adjust the error count for differences in essay length, an error rate for each paper was computed. The error rate was the number of errors divided by the number of words written, the quotient

being multiplied by 1000. The error rate then represents the number of errors per thousand words.

N. Computing the Final Score

For each paper, the final score was the result of the following three elements:

1. a basic mark of 155 points
2. positive points awarded by the marker, up to maximum of 145 points
3. the error rate deducted from the sum of the basic mark and the awarded points

The basic mark of 155 points was given to every paper to avoid the possibility of negative scores after the error rate had been deducted. From the total positive score was deducted the error rate, yielding a final score.

The English Expression Tests were scored by hand according to directions published by Educational Testing Service, and the student's name replaced by a student number to avoid the possibility of contamination.

III. RESULTS

For statistical purposes, the following null hypotheses were tested:

1. There is no statistically significant correlation between scores on the same papers arrived at independently by two raters trained in an objective scoring procedure, especially developed for this experiment.

2. There is no statistically significant difference in the writing achievement of freshmen who have been taught under the experimental method (Guided Revision and Delayed Grades) and similar freshmen who have been taught under the control method (Incidental Revision and Immediate Grades).

3. There are no statistically significant differences among the classes of the five teachers.

4. There is no statistically significant interaction between the class/teacher variable and the method variable.

To test hypothesis 1, above, a reliability coefficient was calculated between scores on the Essay Test independently arrived at by each of the two raters.

To test hypotheses 2, 3, and 4, scores on the English Expression Tests and averaged scores on the Essay Tests were analyzed by means of a two factor (Method and Teacher/Class) Analysis of Covariance, pre-test/post-test design, using the pretest score as the covariate. The results will be discussed for the Essay tests, and then for the English Expression Tests.

The Essay Test

During the second regularly scheduled meeting of the class, all students in the experiment wrote, as a pretest, and, during the regularly scheduled final examination period, as a post-test, an impromptu theme on the general topic of "My Opinion." The time limit for the theme was fifty minutes. Directions for writing the essay were duplicated, distributed, and read to all students by the instructor. The directions suggested topics, possible methods of handling topics, and specific hints as to factors to be considered by the readers in scoring. Attention was especially drawn to the topic statement, to diction, to sentence structure, and to the conclusion.

Interrater reliability

The correlation coefficient (r) between scores computed by Rater A and Rater B was .90 for the pretest and .89 for the post-test. For 100 cases, these correlations are highly significant, and the great care taken in developing the scale and objectifying the scoring procedures proved fruitful. For purposes of the analysis of covariance, the two final scores on each paper were averaged.

Difference between methods

For the essay tests there was no significant difference between the two methods (E and C). The following table will illustrate mean scores for each class in the Experimental group and the Control group on the pretest, the post-test, and the post-test adjusted for differences in the pretest. The scores in the first column represent the entire E and C groups, and within the five subsequent columns the scores for individual classes. The statistical methods for adjusting post-test scores follow Cochran and Cox (1951).

TABLE 1

Mean Scores for Classes and Groups (Essay Test)

		I	II	III	IV	V
PreTest						
XE 157.0	E	154.7	159.8	129.2	152.1	189.3
XC 146.9	C	150.8	144.8	131.9	157.9	149.1
PostTest						
YE 175.0	E	172.2	181.6	160.7	176.4	184.3
YC 172.0	C	178.2	171.4	167.2	178.7	168.1
Adjusted PostTest						
YE 142.65	E	154.66	131.43	306.36	175.5	-54.68
YC 205.1	C	185.62	217.22	295.58	140.68	186.40

TABLE 2

ANALYSIS OF COVARIANCE FOR A TWO-FACTOR RANDOMIZED DESIGN (ESSAY TEST)

SOURCE	SSX	SSP	SSY	DF	SS'Y	MS'Y	F VALUE
Teachers/ Classes	-1420000	-1610000	-1830000	4	60900	15220	15.61**
Methods	1040	1360	18	1	621	621	.637
Interaction	1440000	1620000	1830000	4	54200	13550	13.00**
Error	210000	135000	174000	89	86700	974.6	
Total	235000	145000	179000	98			

**Significant beyond .001 level of confidence

Differences among teachers/classes

Although teachers were chosen so as to be approximately equivalent in ability (all have at least five years' experience in teaching composition and hold at least a Master's degree), and the total Experimental and the total Control groups were balanced as to time of day, and no systematic variability was allowed as to class make-up (announcement of which teachers were to teach which sections was not made at the time students registered themselves in sections), and all teachers assiduously attempted to maintain consistency in all aspects of their teaching other than the difference in handling of papers, Level one differences (among teachers/classes) was significant beyond .001 level. This suggests that the classes may not have been statistically equivalent in writing ability at the pretest, or that the teachers were inadvertantly differing from each other in ways that resulted in statistically significant differences.

Interaction (Teacher/classes by method)

The interactive effect between level one variable and level two variable is also highly significant, beyond the .001 level. Whatever differences that occurred from class to class with each teacher interacted strongly with the difference between the two methods. Examination of the pretest/post-test data reveals that Group V E was considerably higher in writing skills than the other groups, and that the direction of difference between the pretest and post-test changes from group to group. The slight decrease in score, for Group V E, for example, results in a rather astonishing negative score when the post-test is adjusted for the difference from the pretest mean. On the other hand, Group III E, which was considerably below the pretest mean, in adjustment rises above the total possible score. Clearly the interactive effects, which are impossible to determine, are stronger than any difference which can be attributed to the methods. Possibly, the time of day when the post-test essay was scheduled had an important influence on the fluctuation of scores, but this was beyond the control of this investigator.

The English Expression Tests

All students in the experiment took, as a pretest, during the third class meeting, the English Expression Tests of the Cooperative English Tests, published by Educational Testing Service of Princeton, New Jersey. Form 1A was used as pretest, and IB as the post-test, which was given as the second half of the Final Examination given at the end of the semester. The results of the analysis of Covariance follows:

Difference between methods

On the English Expression Tests there was a difference between the total Experimental and the total Control groups significant at the .05 level. This difference, for post-test scores adjusted for

differences from the mean pretest score, was in favor of the Control group. The following table will illustrate mean scores for each class in the Experimental and the Control Group on the pretest, the post-test, and the post-test adjusted for differences in the pretest. The scores in the first column represent the entire E and C groups, and within the five subsequent columns the scores for individual classes.

The statistical methods for adjusting post-test scores follow Cochran and Cox.

TABLE 3

Mean Scores for Classes and Groups (English Expression Tests)

PreTest		I	II	III	IV	V
XE 40.30	E	39.1	43.6	38.7	33.4	46.7
XC 37.34	C	42.0	44.7	28.9	35.7	35.4
PostTest						
YE 35.16	E	39.8	38.3	21.0	33.8	42.9
YC 38.82	C	40.2	41.9	39.6	34.1	38.3
Adjusted PostTest						
YE 24.5	E	38.01	7.71	21.77	62.09	-7.53
YC 48.29	C	19.85	4.27	103.09	54.07	60.19

Difference Significance .05

TABLE 4

ANALYSIS OF COVARIANCE FOR A TWO-FACTOR RANDOMIZED DESIGN (ENGLISH EXPRESSION TEST)

SOURCE	SEX	SSP	SSY	DF	SS'Y	MS'Y	F Value
Teachers/ Classes	-90500	-91500	-91300	4	2680	668.8	7.33**
Methods	467	227	110	1	558	557.7	6.11*
Interaction	92800	92700	94800	4	3370	841.9	9.22**
Error	8910	5660	11700	89	8120	91.25	
TOTAL	11700	6570	15300	98			

* Significant beyond .05 level of confidence

** Significant beyond .001 level of confidence

Differences among Teachers/Classes

As for the Essay Tests, differences among the five different classes of each of the two groups were highly significant (somewhat beyond the .001 level). Since on the objective test, initial differences among the groups appear to be non-significant, it would seem that the continuous process of revision and the time necessarily spent in class to guide such revision had a differential effect on the different classes. Examination of the scores reveals that gains from pretest to post-test seem to change in direction from class to class. Whether this was the result of biases in the teaching or different predispositions of the classes is difficult to surmise from the results.

Interaction (Teacher/classes plus method)

Also similar to the Essay Tests, for the English Expression Tests the interactive effects of the two levels of variability was highly significant, beyond the .001 level. Whatever instruction teachers provided in skills which were measured by this test, such instruction appeared to interact with the main effects of the two methods of teaching composition; or, perhaps, the students' existing state of knowledge in areas such as selecting appropriate vocabulary and being able to select a line in which appeared an error tended to interact with the instructional procedure of consistent revision of errors. In either event, interactive variability was greater than for level one or level two variability.

IV. CONCLUSIONS

The following conclusions can be drawn from this study:

1. Perhaps the most important finding was that, despite the huge body of evidence as to the extreme difficulty in attaining agreement among readers as to the relative quality of samples of student writing, careful preparation of score sheets, a methodical procedure for scoring, and extensive training of the raters can result in high inter-rater reliability.

2. In writing essays, at least under test conditions, students whose revisions were guided by the teachers and whose work was not graded until such revisions were completed achieved no better than students whose work was immediately graded and revisions treated incidentally. However, it may be noted that the adjusted post-test scores reflect rather sharp differences in pretest scores; therefore, the resulting lack of significance may be misleading. Removing the most divergent groups would probably yield an entirely different result. Selecting intact classes for the purpose of the experiment, although administratively the only feasible method at

this University, seems, for this study, to have yielded groups whose original similarity in writing skill is open to question. Judging the effects of revising and rewriting themes that were written at home without the pressure of time, and revised under the same conditions, by the results of an essay written in class under the pressure of test conditions, may be also questionable.

3. The significant difference in favor of the Control Group on the English Expression Tests provides further evidence for the maxim that students will learn directly that which is taught. During the time that the Experimental groups were receiving specific help in revising their themes, Control groups were doing exercises in the Words and Ideas, and probably devoting attention to such matters as were specifically tested by the English Expression Tests. However, this greater knowledge of matters of diction and ability to spot writing errors seem not to have fully carried over into direct writing experiences, as measured by the Essay Tests.

V. RECOMMENDATIONS

Based upon the above stated conclusions, the following recommendations may be offered:

1. Before undertaking any research on the teaching of composition which depends upon scores on an Essay test, careful training of the raters and amendment of the scale being used for those raters and the particular assignment being scored must be undertaken.

2. Random assignment of students to classes, classes to teachers, and groups to experimental conditions would appear to be highly important. Since, for the population of college freshmen, this is frequently extremely difficult due to the administrative procedures of registration, research on the teaching of composition for that population would seem severely hampered. What seems to be needed, rather than further studies of gross curricular matters with a large number of students, using a limited sample of writing, would be small limited studies of carefully circumscribed issues using a small number of students, but based upon a large sample of writing from each subject.

3. The results of an experiment in written composition should, perhaps, be determined by analysis of writing samples gathered over a period of time during the course of the investigation, and should be written under normal instructional conditions. That is, if all themes are written outside of class, then it is these themes which should be subjected to analysis. If all themes are written during the duration of a class period, then the normally written in-class themes should be subjected to analysis.

APPENDIX

Paper Number_____

Student Number_____

SCORE SHEET TO BE USED BY RATING COMMITTEE

A. MARKS AWARDED:	Max.	Student
1. Basic Mark (Give every paper)_____	155	155
2. Material:		
Significance:		
a. Awareness of human issues: political, social, religious, current, or personal.	15	
b. Evidence of Critical Thinking: (defining terms; concrete examples, explaining generalizations; providing evidence, specific instances (2 points each instance)	15	
3. Organization:		
a. Title: (interest and appropriateness)_____	5	
b. Introduction: Thesis statement and interest arousal	10	
c. Logical sequence of paragraphs: Order in which paragraphs appear. Divide 10 pts. by number of paragraphs and award according to merit.	10	
d. Unity within Paragraphs: Relevancy of material within paragraphs. (Score same as c.)	10	
e. Transition between Paragraphs: phrases, clauses, transitional words (score same as c)	10	
f. Effective Conclusion: Summation and closing remarks	10	
4. Sentences:		
a. Variety in sentence structure: Complex, compound, various kinds of subordination adjective, adv. clauses, verbals. (2 points per instance)	10	
b. Grammatical correctness, lucidity	10	
c. Effective predication (Guth Ch. 13) (2 pts. each instance)	10	
5. Diction: and Vocabulary		
a. Exactness and vividness of nouns, verbs, adjectives, etc. (1 pt. for each noteworthy choice)	10	
b. Interesting and appropriate figures of speech, comparisons, illustrations, allusions, quotations (2 pts. per instance)	10	
c. Originality of expression (use of humor; exaggeration for effect; mock seriousness; anticlimax; understatement; pretentious language used for effect; etc.) 2 pts. each instance	10	
	300	

APPENDIX

Paper Number_____

Score Sheet (continued)

B. DEDUCTIONS:

Spelling	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Punctuation	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Usage	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Grammar	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Sentence	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Form	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
TOTAL:	

C. ERROR RATE:

Length:_____words. Number of errors_____

Error Rate_____ x 1,000 _____.

List of References

Borg, Walter R. Educational Research: An Introduction. New York: David Mckay, 1963.

Braddock, Richard; Lloyd-Jones, Richard; and Schoer, Lowell. Research in Written Composition. Champaign, Illinois: National Council of Teachers of English, 1963.

Braddock, Richard. "Written Composition" in Encyclopedia of Educational Research edited by Robert L. Ebel. New York: MacMillan, 1969.

Burton, Dwight L. and Arnold, Lois V. Effects of Frequency of Writing and Intensity of Evaluation upon High School Students' Performance in Written Composition. United States Office of Education Cooperative Research Project 1523, Tallahasee: Florida State University, 1963.

Buxton, Earl W. "An Experiment to Test the Effects of Writing Frequency and Guided Practice Upon Students' Skill in Written Expression." Unpublished Ph.D. Dissertation, Stanford University, 1958.

Campbell, Donald T. and Stanley, J.C. "Experimental and Quasi Experimental Designs for Research in Teaching," in Handbook of Research in Teaching, edited by N.L. Gage, Chicago: Rand McNally, 1963.

Cochran, William C., and Cox, Gertrude M. Experimental Designs. New York: John Wiley & Sons, 1951.

Gorrell, Robert. "Freshman Composition," in College Teaching of English, edited by John C. Gerber. New York: Appleton-Century-Crofts, 1965.

Manual for Interpreting Scores. Cooperative English Tests. Princeton, N.J.: Educational Testing Service, 1960.

Pierson, Howard. "Peer and Teacher Correction: A Comparison of the Effects of Two Methods of Teaching Composition in Grade Nine Classes." Unpublished Doctoral Dissertation, New York University, 1967

Seidman, Earl. "Marking Students' Compositions: Implications of Achievement Motivation Theory." Dissertation Abstracts, 1968, 28, 2605.

Sherwin, J. Stephen. Four Problems in Teaching English: A critique of Research. Scranton: International Textbook, 1969.

Skinner, B.F. Cumulative Record. New York; McMillan, 1953.

Sutton, Joseph T. and Allen, Elliott D. "The Effect of Practice and Evaluation on Improvement in Written Composition." Cooperative Research Project 1993, Stetson University, 1964.

Zoellner, Robert. "A Behavioral Approach to Writing," College English, Volume 30 (January 1969).